

# Advanced Topics in Data Science

**2019-2020 Academic Year**  
**Master of Research in Economics, Finance and Management**

## 1. Description of the subject

- Advanced Topics in Data Science
  - Total credits: 6 ECTS
  - Type of subject: Elective
  - Department of Economics and Business
  - Teaching team: David Rossell
- Code: 32282  
Workload: 150 hours  
Term: 1st

## 2. Teaching guide

### 🔗 Introduction

Statistical and Machine learning techniques are having a deep effect on many disciplines, including Economics. As a consequence of the increasing number of applications using a combination of data science with techniques popular in Economics (see references) many World-leading institutions have incorporated data science in their PhD programs. As examples see the set of lectures in the NBER Summer Institute of 2015 by Guido Imbens and Susan Athey (in particular lecture 3, [www.nber.org/econometrics\\_minicourse\\_2015/](http://www.nber.org/econometrics_minicourse_2015/)), the Big Data and machine learning syllabus at the Harvard Economics PhD program (<https://locator.tlt.harvard.edu/course/colgsas-156429>), the [Chicago Booth](#) PhD courses on Bayesian inference, Big Data and Machine Learning, or the [Bocconi](#) PhD courses in text analysis, Econometrics of networks and causal analysis.

Statistical methods in data science are rapidly evolving to cope with increasingly challenging problems, a case in point being high-dimensional situations where one considers a large or even infinite number of parameters or models. Engaging effectively in research or cross-disciplinary collaborations requires data scientists to be able to absorb, and when needed modify or extend, such methodology. Just as importantly, they need to communicate such ideas effectively to a potentially non-expert audience.

The goal of this course is to introduce students to some foundations behind these methods, with a certain emphasis on the Bayesian framework, penalized likelihood and latent variable methods (e.g. as used in text analysis) methods, expose them to and discuss research literature, and practice the skills needed for applying and presenting novel research. The learning outcomes are an improved familiarity with selected research topics in Statistics that are relevant for Data Science, at a level sufficient to critically appraise, modify and apply novel methods, and improved oral and written presentation skills. The course also intends to provide students with applied data analysis skills useful for their MRes thesis and subsequent research work.

Pre-requisites: the course is designed for MRes students who are familiar with basic Statistical inference, specifically linear regression and maximum likelihood estimation. Basic R programming skills are beneficial, though examples and links to learning resources will be provided for students who are not familiar with R.

### 🔗 Contents

1. Foundations. We briefly review classical results from maximum likelihood estimation and computational methods such as the bootstrap, and we introduce the basic Bayesian paradigm for statistical inference, its use for model selection, parameter estimation and prediction, and standard computational tools such as Gibbs or Metropolis-Hastings.
2. Foundations of variable selection in high-dimensional regression. We review the fundamental penalized likelihood and Bayesian frameworks for linear regression models with a large number of variables. We shall discuss the relative merits of current strategies such as LASSO, adaptive LASSO and related penalties to help decouple variable selection from prediction or non-local priors and other Bayesian strategies to achieve good performance in high dimensions. We will discuss theoretical and practical considerations, including strategies to carry out computations in a scalable manner.
3. Mixture models and Bayesian non-parametrics. We shall move towards more flexible models enabled by the use of mixtures. We will consider finite and infinite mixture models, specifically for the latter we shall focus on their use as the basic building block for a broad family of Bayesian non-parametric methods. We shall discuss applications to mixture-of-regressions models to account for unobserved confounders that may bias inference, to hidden Markov models for time series and to text data analysis.

4. Beyond linear regression. We will overview selected strategies to extend the linear regression framework, such as generalized linear models, flexible models for count data, random forests and Bayesian additive regression trees, or capturing non-linear relationships.

**References.** The books below provide a good introduction to a substantial part of the topics covered in this course (and many others), however we shall complement them with a number of additional selected research manuscripts.

- Andrew Gelman, John B. Carlin, Hals S. Stern, David B. Dunson, Aki Vehtari, Donald B. Rubin. Bayesian Data Analysis (3rd edition). CRC Press, 2013.
- Trevor Hastie, Robert Tibshirani, Martin Wainwright. Statistical learning with sparsity. The LASSO and its generalizations. CRC press.
- Sara van de Geer, Peter Bühlman. Statistics for high-dimensional data: methods, theory and applications. Springer, 2001.
- Nils Lid Hjort, Chris Holmes, Peter Müller, Stephen G. Walker. Bayesian non- parametrics. Cambridge University Press, 2010.
- Sylvia Frühwirth-Schnatter. Finite mixture and Markov switching models. Springer, 2006.

#### **Some papers on data science and economics:**

- Chernozhukov, V., A. Belloni, C. Hansen (2014), "High-Dimensional Methods and Inference on Treatment and Structural Effects in Economics", J. Economic Perspectives
- Chernozhukov, V., D. Chen, A. Belloni, C. Hansen (2012), "Sparse Models and Methods for Instrumental Regression, with an Application to Eminent Domain", Econometrica
- Einav, L. y J. Levin (2014), "Economics in the age of big data," Science, 346 (6210).
- Varian, H. (2014), "Big data: new tricks for econometrics," Journal of Economic Perspectives, 28 (2), 3-28.

#### **🔗 Teaching methodology**

The course will be delivered in a combination of regular lectures, computer-based seminars where students get hands-on experience with the taught data analysis methods, and presentations by students (on published manuscripts chosen by the students and on the final project).

#### **🔗 Assessment and Grading System**

Students will be asked to orally present 2 research papers of their choice (40% of the final mark), some selected exercises from the seminar sessions (10% of final mark) and a written report (50% of the final mark). This project will be decided by the students but must be pre-approved by the lecturer, and should involve the application, critical assessment or extension of the research methods seen in class. The content can be theoretical, empirical, a practical application or a combination of the former.